# 一般化同値関係上のカバー配列の計算および文脈自由文法のサブクラスの学習に関する研究

修士学位論文要約（令和３年３月）

# 一般化同値関係上のカバー配列の計算および文脈自由文法のサブクラスの学習に関する研究

菊池 なつみ
指導教員：篠原 歩, 学位論文指導教員：吉仲 亮

# Computing Cover Arrays on General Equivalence Relations and Learning a Subclass of Context-free Grammars

Natsumi KIKUCHI
Supervisor: Ayumi SHINOHARA, Research Advisor: Ryo YOSHINAKA

We achieved results in terms of string and formal language regularities, respectively. First, we generalize the notion of covers for SCERs and prove that existing algorithms to compute the shortest cover array and the longest cover array of a string under the identity relation will work for any SCERs taking the accordingly generalized border arrays. Second, we propose a canonical grammar for $k,l$-substitutable languages by extending the canonical grammar defined by Clark for substitutable languages. We show that a subclass of $k,l$-substitutable context-free languages is identifiable with the proposed grammar in the limit from positive data by a exponential-time algorithm.

## 1. はじめに

規則性を発見することは文字列処理のおける重要なタスクであり, パターン照合, データ圧縮, 文法推論などのタスクに応用が期待されている. 規則性を考える対象として文字列と言語に大きく分けて考えることができる.

文字列に関しては, 周期, ボーダー, カバーなどの多くの規則性の研究が行われている. 周期とボーダーについては, 恒等関係上に留まらず様々な同値関係で研究がなされ, 近年では *substring consistent equivalence relation (SCER)* という同値関係のサブクラスに対しても研究が行われた [5]. 本研究では, SCER 上でのカバーの計算について研究を行う.

形式言語に関しては, 規則性を文法として獲得する文法推論の研究が盛んに行われている. 文脈自由言語は正規言語よりも表現力が高くより自然言語的である一方, 学習は困難であることが知られている. 近年では, 文脈自由言語のサブクラスに対象を絞った置換可能言語 [3] やその一般化である $k,l$-置換可能言語 [6] について研究が行われ, 効率の良い言語レベルの学習が可能であることがわかった. 特に, 置換可能言語については言語を一意に表現する標準文法が提案され, その文法を用いて効率の良い文法レベルの学習が可能であることがわかっている [2]. 本研究では, $k,l$-置換可能言語に対しても同様に結果が得られるかについて研究を行う.

## 2. 一般化同値関係上のカバー配列の計算

文字列 $T$ の全ての位置が文字列 $C$ と等価な $T$ の部分文字列内にある時, $C$ は $T$ のカバーである. 本研究では, このカバーの概念を次のように定義される SCER 上に対して $\approx$-カバーとして一般化する. 同値関係 $\approx$ について, 任意の文字列 $X, Y \in \Sigma^*$ に対して, $X \approx Y$ ならば (1) $|X| = |Y|$, そして (2) 任意の $1 \leq i \leq j \leq |X|$ について $X[i:j] \approx Y[i:j]$ が成り立つとき, $\approx$ を SCER であるという.

$T$ の各接頭辞の最短カバー, および最長カバーの長さを表す配列は, それぞれ $T$ の最短カバー配列, 最長カバー配列と呼ばれ, 入力である $T$ のボーダー配列を用いて線形時間で求めるアルゴリズムが既に存在している [1)4)]. 本研究では, 入力を $\approx$-ボーダー配列に置き換えることで, それらのアルゴリズムが SCER 上においても同様に動作することを証明し, 次の定理が得られた.

**定理 1** 長さ $n$ の文字列 $T$ の $\approx$-ボーダー配列が与えられたとき, *Breslauer* のアルゴリズム [1] を用いることで, $O(n)$ 時間で $T$ の最短 $\approx$-カバー配列を求めることができる.

**定理 2** 長さ $n$ の文字列 $T$ の $\approx$-ボーダー配列が与えられたとき, *Li & Symth* のアルゴリズム [4] を用いることで, $O(n)$ 時間で $T$ の最長 $\approx$-カバー配列を求めることができる.

また, 最長 $\approx$-カバー配列を求めるアルゴリズムについては, 既存アルゴリズムの簡略版も提案した.

## 3. 文脈自由文法のサブクラスの学習

まず, Clark が置換可能言語に対して定めた標準文法 [2] を拡張し, 次のように定義される $k,l$-置換可能言語に対して標準文法を提案する. 言語 $L$ に対し, 非負整数の長さ $k,l$ の任意の $u,v$ において, $x_1uy_1vz_1, x_1uy_2vz_1, x_2uy_1vz_2 \in L \Rightarrow x_2uy_2vz_2 \in L$ が成り立つとき, $L$ は $k,l$-置換可能言語であるという. $0,0$-置換可能言語は置換可能言語である. 厳密には, 次に定義する $k,l$-置換可能言語のサブクラスに対して標準文法を定める.

**定義 1** 言語クラス $\mathcal{L}_{k,l\text{-}SC}$ を次の条件を満たす言語 $L$ の集合とする.

1. $\$^kL\$^l$ は $k,l$-置換可能言語である. ここで, \$ はダミー記号である.
2. 空言語でなく, 空文字列を含まない言語である.
3. 有限個の素な合同類を持つ.

三つ目の条件によって, $\mathcal{L}_{k,l\text{-}SC}$ は文脈自由言語のサブクラスとなる. $\mathcal{L}_{0,0\text{-}SC}$ は Clark が標準文法を定めた言語クラスと一致する.

**定義 2 ($k,l$-標準文法)** 言語 $L \in \mathcal{L}_{k,l\text{-}SC}$ に対して, 次の条件を満たす文脈自由文法 $G_*(L) = \langle \Sigma, V, P, S\rangle$ を $L$ の $k,l$-標準文法という. 非終端記号集合 $V$ を言語 $L$ の持つ素な合同類と開始記号 $S$ の集合とする. 生成規則集合 $P$ は次の生成規則の集合とする.

1. $S$ を左辺に持つ, $L$ の素分解の生成規則. $L$ の任意の素分解 $X_1 \dots X_n$ に対し, $S \to X_1 \dots X_n$.
2. $L$ に現れる終端記号 $a \in \Sigma$ の生成規則 $[a] \to a$.
3. $L$ の全ての有効な生成規則.

$\mathcal{L}_{0,0\text{-}SC}$ の持つ合同類が一意な素分解を持つのに対し, $\mathcal{L}_{k,l\text{-}SC}$ の合同類は複数の素分解を持つ場合がある. 合同類の最短要素の長さの指数通りの素分解を持つ言語が存在する. Clark が定めた標準文法の生成規則に対し素分解のバリエーションを考慮したものが $k,l$-標準文法である.

次に, $k,l$-標準文法を用いた言語クラス $\mathcal{L}_{k,l\text{-}SC}$ の学習の概略を Algorithm 1 に示す. 関数 WeakLearner は, 言語レベルの学習を行う関数 [6] であり, 文字列集合を入力とし, 目標言語を表す文法を出力する. 関数 Part は, 文法と文字列集合を入力とし, 入力の部分文字列集合の合同類分割を出力する. 関数 Prime は, 合同類集合を入力とし, その中の素な合同類の集合を出力する. 関数 Decomp は, 合同類を入力とし, その全ての素分解の集合を出力する. また, $S$ は開始記号である. Algorithm 1 について次の定理が成り立つ.

**Algorithm 1:** $\mathcal{L}_{k,l\text{-}SC}$ の文法レベルの学習

**Input** : 文字列の列 $w_1, w_2, \dots$
**Output:** 文脈自由文法の列 $G_1, G_2, \dots$

**for** $n = 1, 2, \dots$ **do**
  $D = \{w_1, \dots, w_n\}$;
  $\hat{G} = \mathsf{WeakLearner}(D)$;
  $\mathcal{C} = \mathsf{Part}(\hat{G}, D)$;
  $Pr = \mathsf{Prime}(\mathcal{C})$;
  $P_{\mathrm{L}} = \{C \to a \mid C \in \mathcal{C}, a \in C\}$;
  $\mathcal{C}' = \{C \in \mathcal{C} \mid C \cap D \neq \emptyset\}$;
  $P_{\mathrm{I}} = \{S \to \alpha \mid C \in \mathcal{C}', \alpha \in \mathsf{Decomp}(C)\}$;
  $P_{\mathrm{B}} = \emptyset$;
  **for** $N, M \in Pr, C \in \mathcal{C}$ **do**
    **for** $\alpha \in \mathsf{Decomp}(C)$ **do**
      $R = (N \to M\alpha)$;
      **if** $R$ が有効である **then**
        $P_{\mathrm{B}} \leftarrow P_{\mathrm{B}} \cup \{R\}$;
  $G_n = \langle \Sigma, Pr \cup \{S\}, P_{\mathrm{L}} \cup P_{\mathrm{B}} \cup P_{\mathrm{I}}, S\rangle$;
  output $G_n$;

**定理 3** 言語クラス $\mathcal{L}_{k,l\text{-}SC}$ は, 指数時間の更新を行う *Algorithm 1* によって正例から文法レベルの学習が可能である.

## 文献


1) D. Breslauer. An on-line string superprimitivity test. *Inform. Process. Lett.*, Vol. 44, No. 6, pp. 345–347, 1992.
2) A. Clark. Learning trees from strings: A strong learning algorithm for some context-free grammars. *J. Mach. Learn. Res.*, Vol. 14, No. 1, pp. 3537–3559, 2013.
3) A. Clark and R. Eyraud. Polynomial identification in the limit of substitutable context-free languages. *J. Mach. Learn. Res.*, Vol. 8, pp. 1725–1745, 2007.
4) Y. Li and W. F. Smyth. Computing the Cover Array in Linear Time. *Algorithmica*, Vol. 32, No. 1, pp. 95–106, 2002.
5) et al Matsuoka. Generalized pattern matching and periodicity under substring consistent equivalence relations. *Theor. Comput. Sci.*, Vol. 656, pp. 225–233, 2016.
6) R. Yoshinaka. Identification in the limit of $k,l$-substitutable context-free languages. In *ICGI*, pp. 266–279. Springer, 2008.